Finishing and Special Motifs: Lessons Learned From CRISPR Analysis Using Next Generation Draft Sequences

7th Annual Sequencing, Finishing, and Analysis in the Future Meeting

Catherine E. Campbell, PhD

June 8, 2012

noblis

For the best of reasons

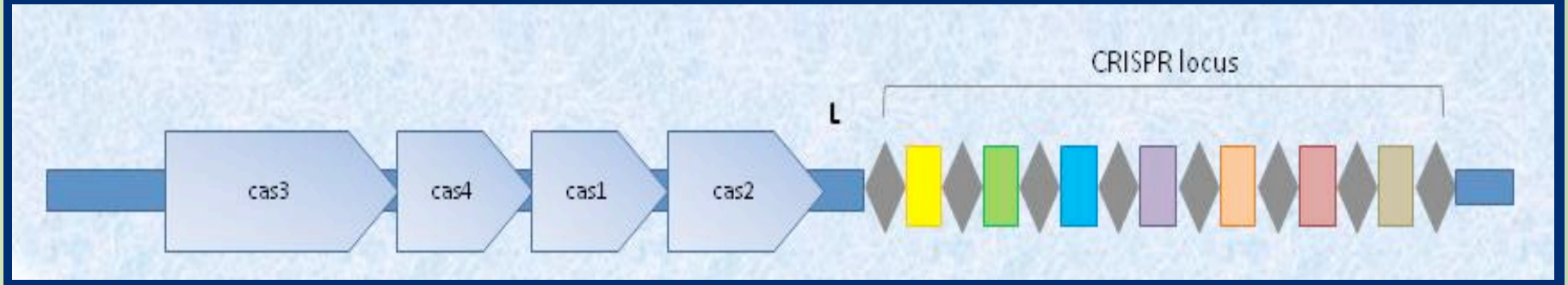
CRISPR and Assembly

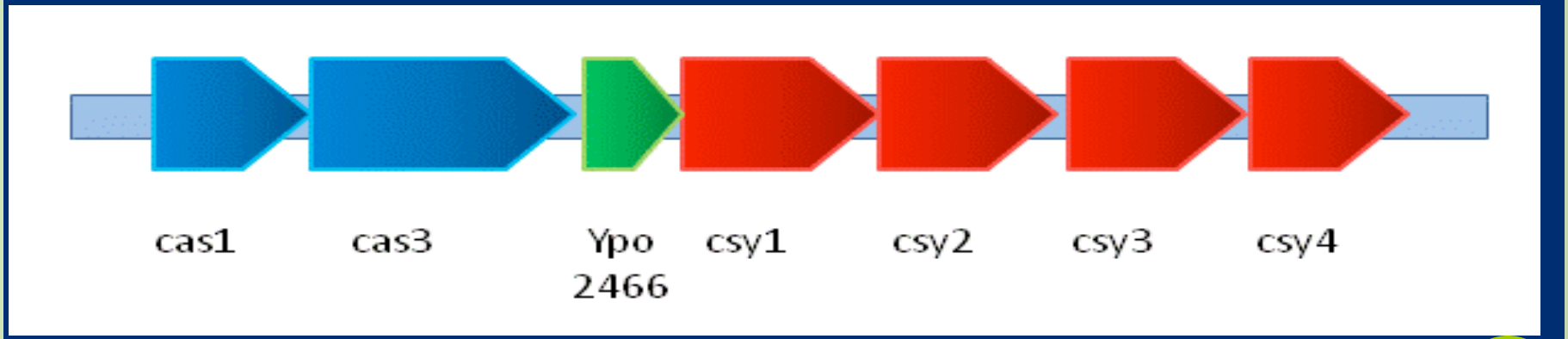
- What are CRISPR?
- What is their structure and function?
- WGS Assembly and CRISPR
- How can CRISPR be identified from unassembled reads?
- Spacer mapping
- Spacer anomalies
- Using CRISPR for phylogeny
- Horizontal gene transfer

What Are CRISPR?

- Clustered Regularly Interspaced Short Palindromic Repeats
- They are a family of repetitive DNA sequences that are separated by short, generally unique DNA sequences.
- About 40% of sequenced Bacteria and 90% of sequenced Archaea have CRISPR systems
- CRISPR may act as an adaptive immune system against invading phages and plasmids

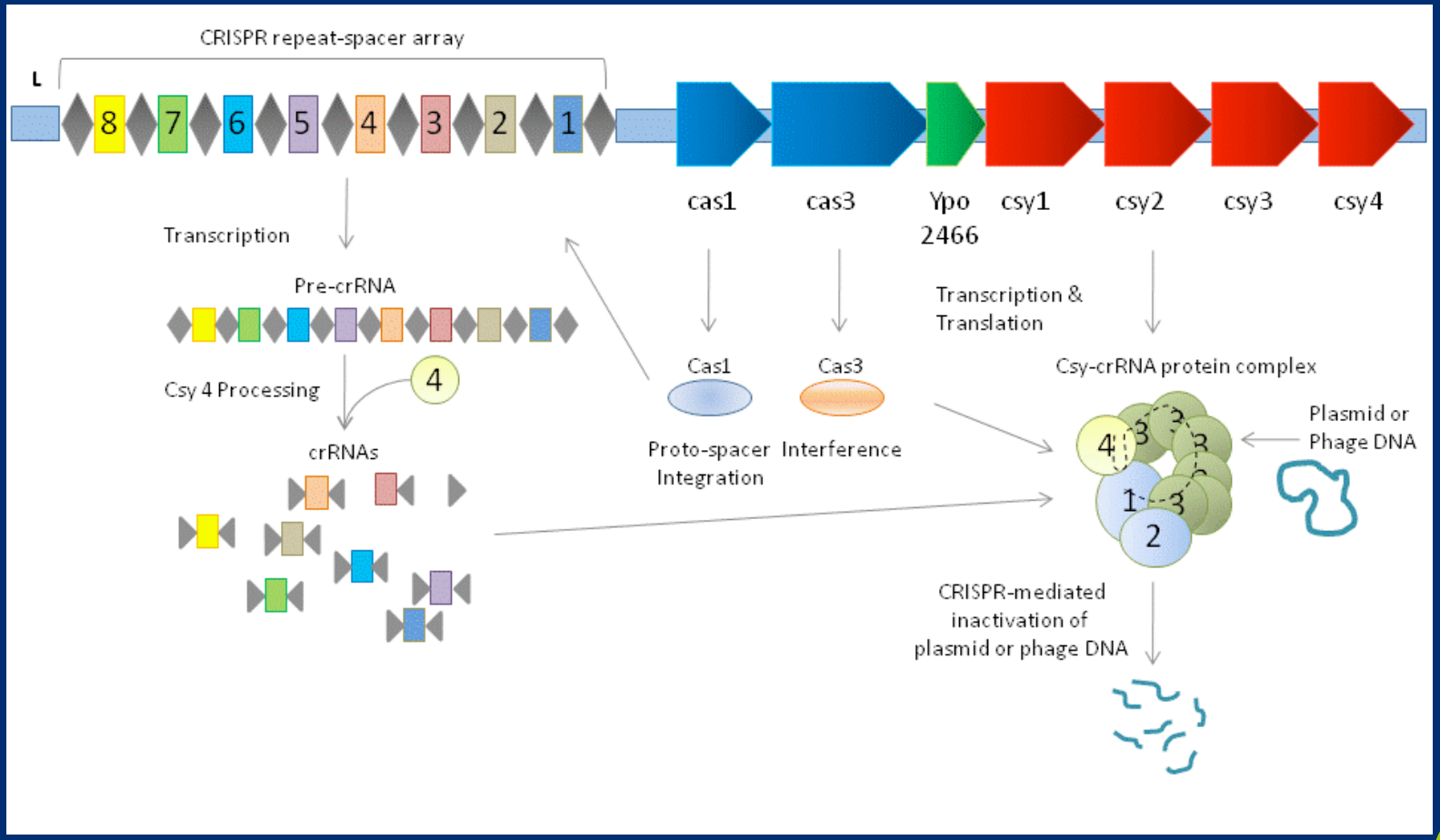
CRISPR STRUCTURE





noblis
For the best of reasons

Mechanism of Action



noblis
For the best of reasons

Strains Used

Source	Location	Year Isolated
Mountain Lion	Arizona	2007
Soil	Arizona	2007
Human	Arizona	2007
Human	New Mexico	2002
Mouse Passage	New Mexico	2002
Flea	New Mexico	2002
Flea	New Mexico	2002
Squirrel	Colorado	2007
Squirrel	Colorado	2007
Squirrel	Colorado	2007
Squirrel	Colorado	2007
Rabbit	Colorado	2007
Flea	Colorado	2007

Assembly Methods Overview

De novo Sequence Assembly

Identify CRISPR repeats

Identify spacers and CRISPR loci

Order and map spacers within loci

Group loci based on spacer sequences

Create phylogenetic trees from spacer sequences

Identify *cas* genes

Associate *cas* genes with CRISPR loci

Assembly Methods- Part 1

- The direct repeat sequence (TTTCTAAGCTGCCTGTGCGGCACTGAAC) was used to identify the CRISPR loci.
- NGS reads were assembled using the command line version of the Newbler *de novo* assembler (Roche).
- *cas* genes were identified using BLAST to find the seven genes associated with the Ypest subtype locus
  - each *cas* gene was aligned to the assembled contigs to determine their presence in each genome.
- Repeats were identified using bl2seq (NCBI), which aligned the repeat sequence to each contig and identified all contigs containing repeats.

noblis
For the best of reasons

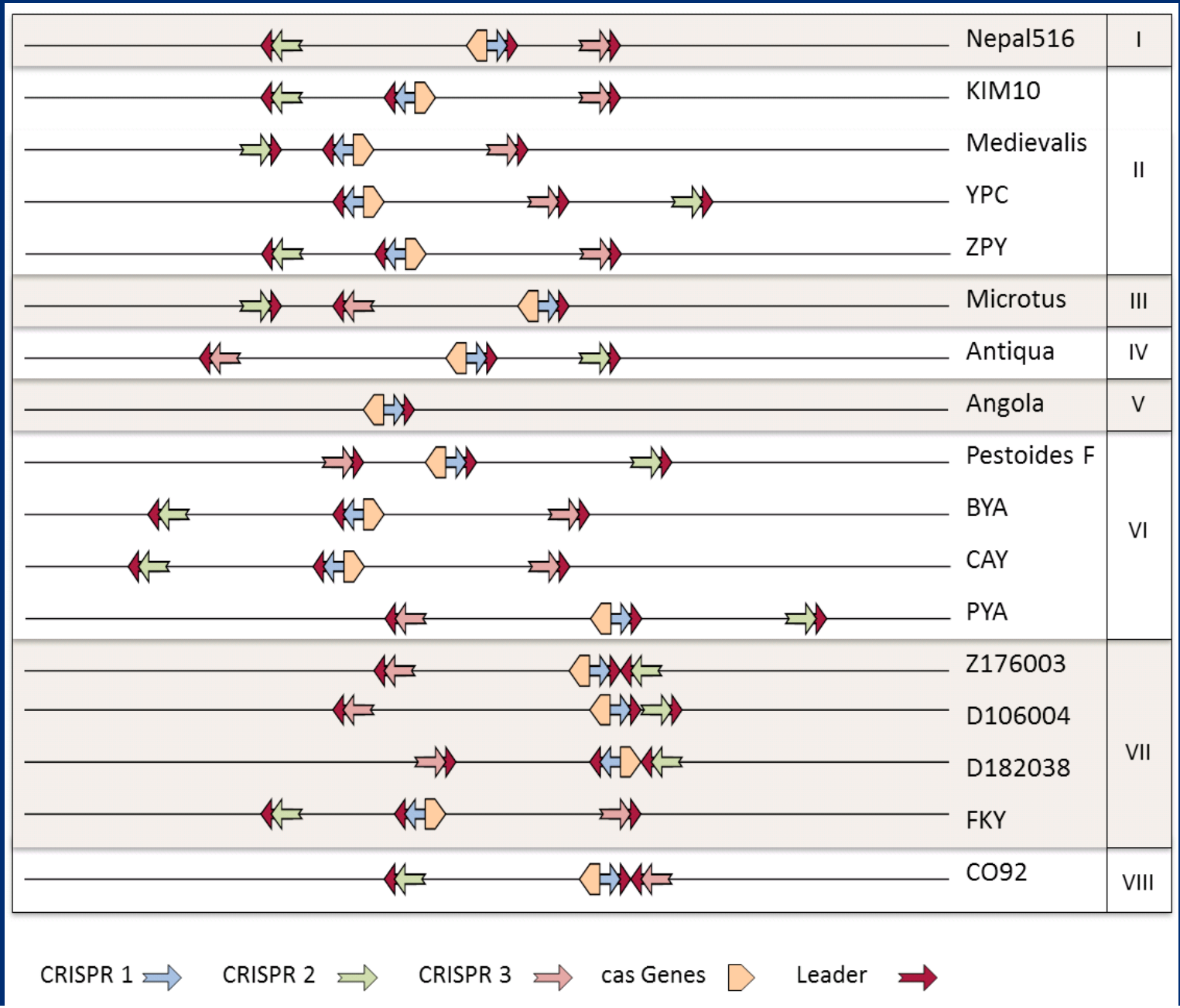
Assembly Methods-Part 2

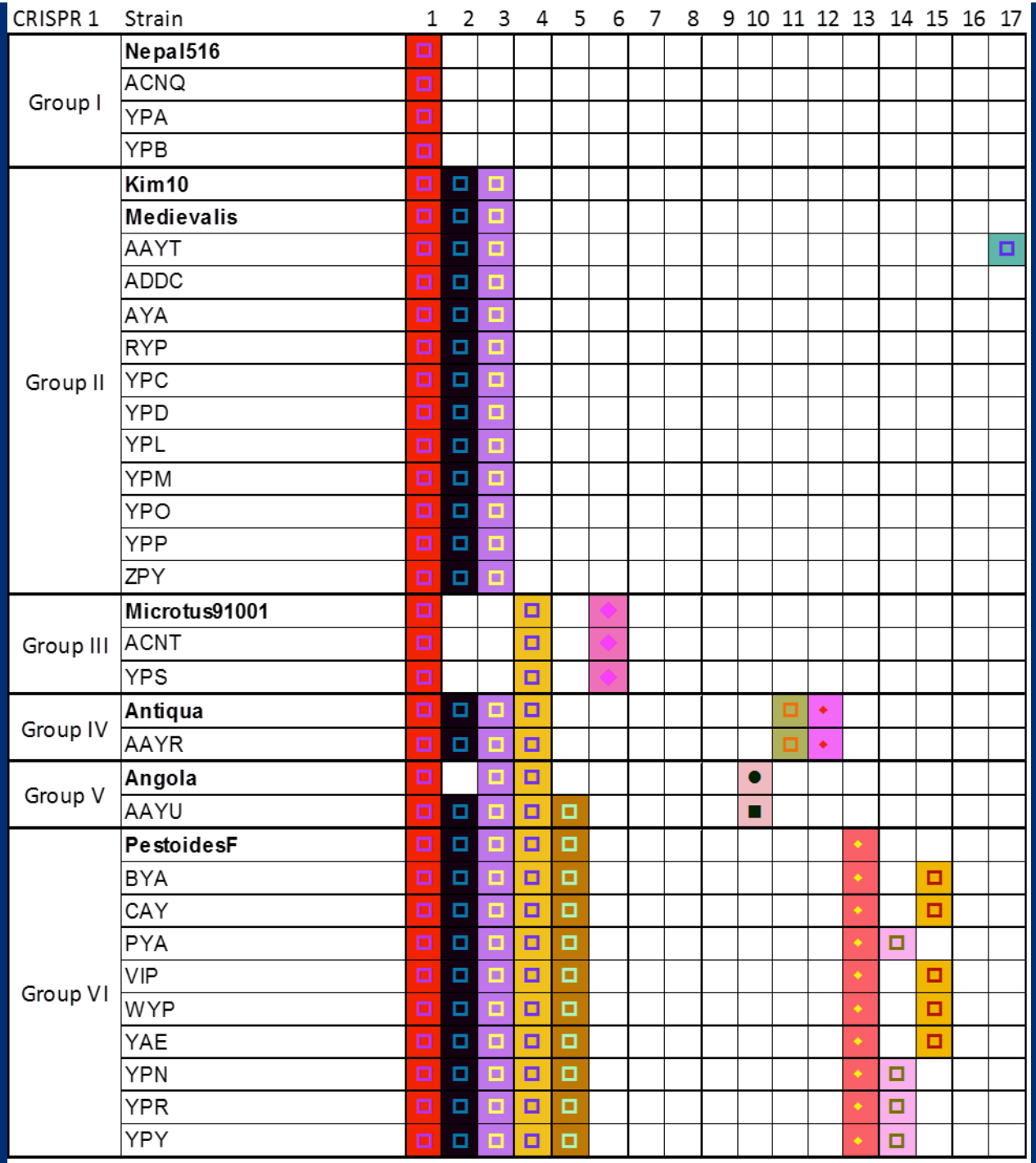
- A Perl script, IDcrispr, parsed the BLAST formatted output of bl2seq by ordering the starting location.
  - The repeat was identified if it occurred within 200bp and on the same strand of an adjacent repeat on the alignment list.
- A second Perl script was used to extract spacers sequences.
- Spacer sequences were input into a spacer visualization tool.
- Spacer sequences were aligned using ClustalW, and the output was used to generate phylogenetic trees.

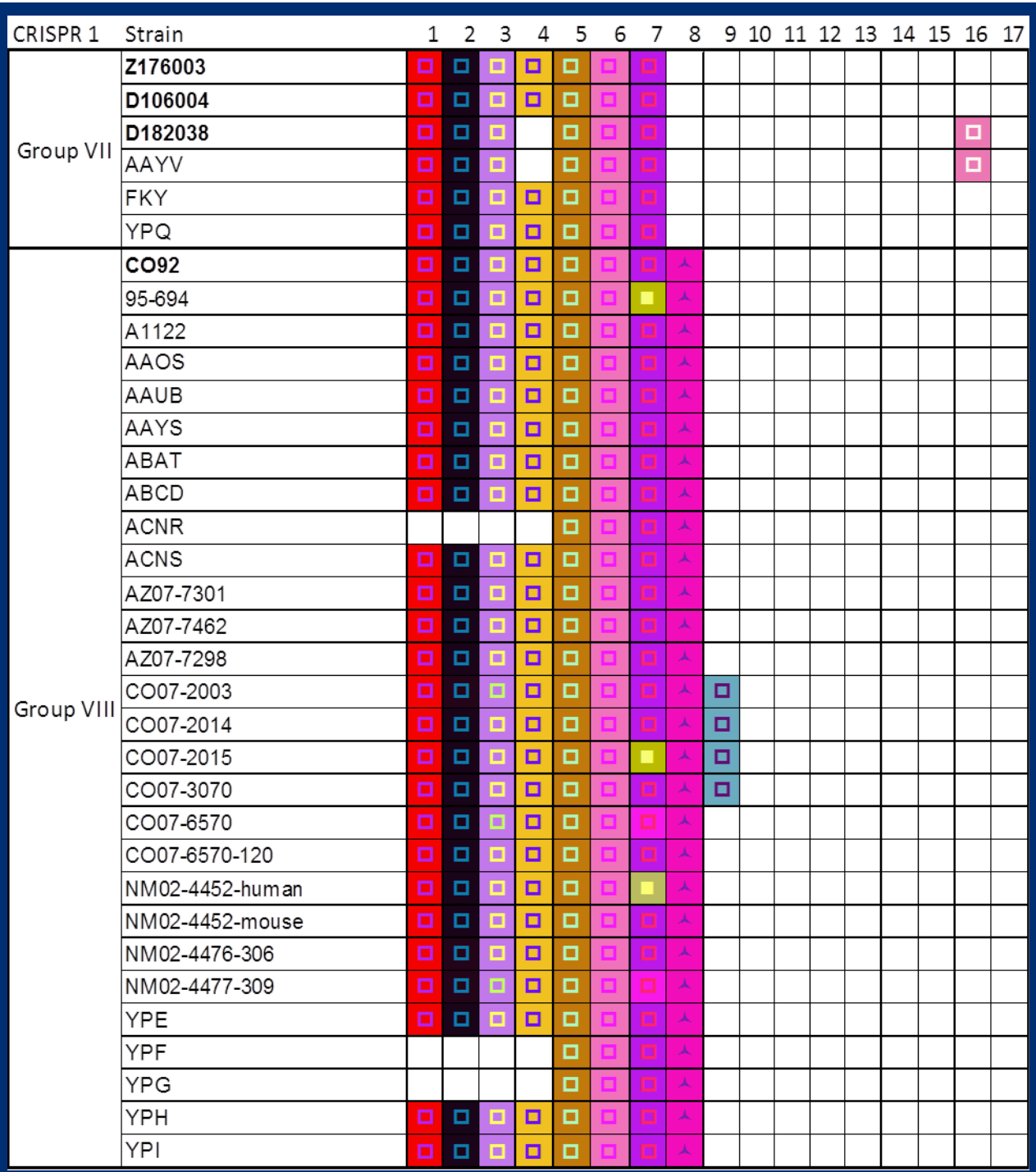
Assembly Methods- Issues Encountered

- Extra or missing spacers were verified by inspecting the raw read data.
  - When an extra spacer(s) was observed in one of the contigs in an assembled genome, BLAST was used to look for that spacer sequence within the unassembled raw reads from all of the other samples.
  - Additionally, the spacer with its flanking repeat sequence was verified by BLAST alignment to identify whether the spacer was not only present, but was also adjacent to a direct repeat.
- In some cases spacers appeared to be missing
  - The repeat-spacer-repeat sequence was aligned to the raw reads using BLAST, to determine if the sequence was truly missing or was present but had not been assembled into the contig.
  - In one sample our identification pipeline failed to identify spacers 5 and 6 in CRISPR1. Further investigation revealed that CRISPR1 spanned two contigs, and the missing spacers were found unassembled in the raw reads, flanked by the CRISPR repeat sequence.

Spacer Mapping in *Y. pestis*




CRISPR 1
Strain
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
Group I
Nepal516
ACNQ
YPA
YPB
Group II
Kim10
Medievalis
AAYT
ADDC
AYA
RYP
YPC
YPD
YPL
YPM
YPO
YPP
ZPY
Group III
Microtus91001
ACNT
YPS
Group IV
Antiqua
AAYR
Group V
Angola
AAYU
Group VI
PestoidesF
BYA
CAY
PYA
VIP
WYP
YAE
YPN
YPR
YPY


CRISPR 1
Strain
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
Group VII
Z176003
D106004
D182038
AAYV
FKY
YPQ
Group VIII
CO92
95-694
A1122
AAOS
AAUB
AAYS
ABAT
ABCD
ACNR
ACNS
AZ07-7301
AZ07-7462
AZ07-7298
CO07-2003
CO07-2014
CO07-2015
CO07-3070
CO07-6570
CO07-6570-120
NM02-4452-human
NM02-4452-mouse
NM02-4476-306
NM02-4477-309
YPE
YPF
YPG
YPH
YPI

CRISPR 2	Strain	1	2	3	4	5	6	7	8	9	10
Group I	**Nepal516**										
	ACNQ										
	YPA										
	YPB										
Group II	**Kim10**										
	Medievalis										
	AAYT										
	ADDC										
	AYA										
	RYP										
	YPC										
	YPD										
	YPL										
	YPM										
	YPO										
	YPP										
	ZPY										
Group III	**Microtus91001**										
	ACNT										
	YPS										
Group IV	**Antiqua**										
	AAYR										
Group V	**Angola***										
	AAYU										
Group VI	**PestoidesF**										
	BYA										
	CAY										
	PYA										
	VIP										
	WYP										
	YAE										
	YPN										
	YPR										
	YPY										

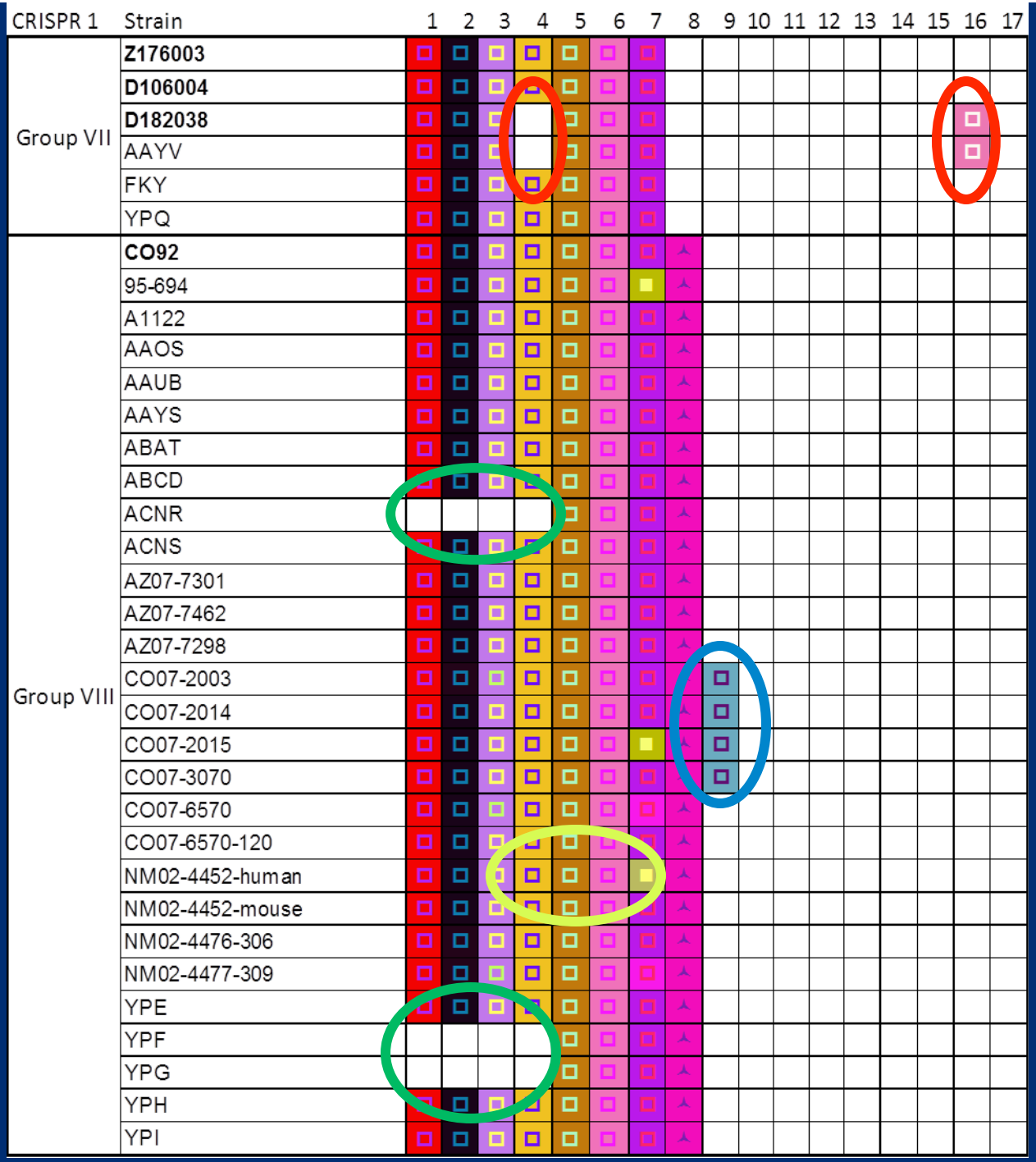
*The Angola strain does not contain CRISPR locus 2 or 3.

CRISPR 2	Strain	1	2	3	4	5	6	7	8	9	10
Group VII	**Z176003**										
	D106004										
	D182038										
	AAYV										
	FKY										
	YPQ										
Group VIII	**CO92**										
	95-694										
	A1122										
	AAOS										
	AAUB										
	AAYS										
	ABAT										
	ABCD										
	ACNR										
	ACNS										
	AZ07-7301										
	AZ07-7462										
	AZ07-7298										
	CO07-2003										
	CO07-2014										
	CO07-2015										
	CO07-3070										
	CO07-6570										
	CO07-6570-120										
	NM02-4452-human										
	NM02-4452-mouse										
	NM02-4476-306										
	NM02-4477-309										
	YPE										
	YPF										
	YPG										
	YPH										
	YPI										

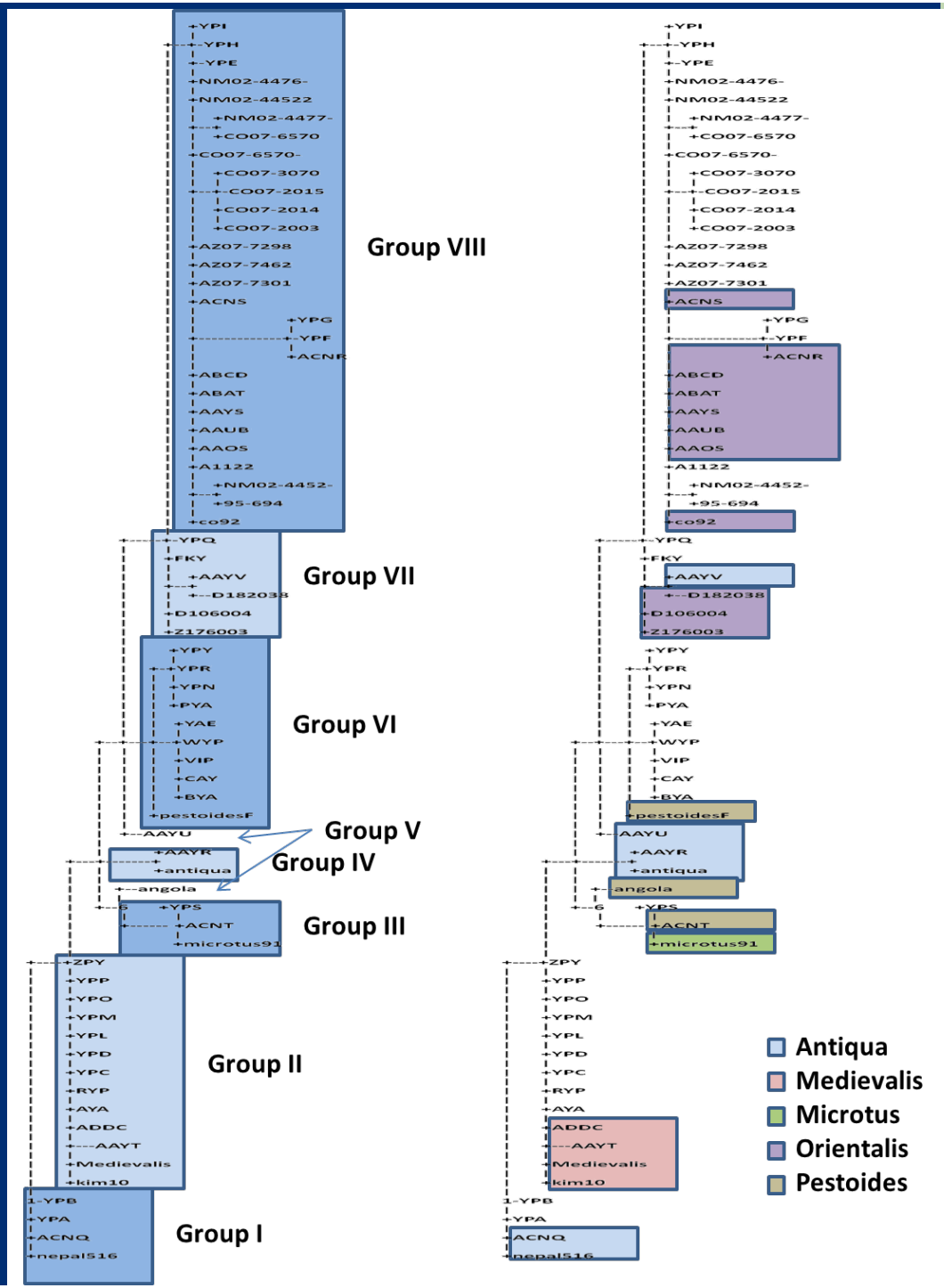
CRISPR 3	Strain	1	2	3	4	5
Group I	**Nepal516**	■	□	□		
	ACNQ	■	□	□		
	YPA	■	□	□		
	YPB	■	□	□		
Group II	**Kim10**	■	□	□		
	Medievalis	■	□	□		
	AAYT	■	□	□		
	ADDC	■	□	□		
	AYA	■	□	□		
	RYP	■	□	□		
	YPC	■	□	□		
	YPD	■	□	□		
	YPL	■	□	□		
	YPM	■	□	□		
	YPO	■	□	□		
	YPP	■	□	□		
	ZPY	■	□	□		
Group III	**Microtus91001**	■				
	ACNT	■	□			
	YPS	■	□			
Group IV	**Antiqua**	■	□	□		
	AAYR	■	□	□		
Group V	**Angola***					
	AAYU	■	□	□		
Group VI	**PestoidesF**	■	□	□	⅄	■
	BYA	■	□	□	⅄	■
	CAY	■	□	□	⅄	■
	PYA	■	□	□	⅄	■
	VIP	■	□	□	⅄	■
	WYP	■	□	□	⅄	■
	YAE	■	□	□	⅄	■
	YPN	■	□	□	⅄	■
	YPR	■	□	□	⅄	■
	YPY	■	□	□	⅄	■

*The Angola strain does not contain CRISPR locus 2 or 3.

CRISPR 3	Strain	1	2	3	4	5
Group VII	**Z176003**	■	□	□		
	D106004	■	□	□		
	D182038	■	□	□		
	AAYV	■	□	□		
	FKY	■	□	□		
	YPQ	■	□	□		
Group VIII	**CO92**	■	□	□		
	95-694	■	□	□		
	A1122	■	□	□		
	AAOS	■	□	□		
	AAUB	■	□	□		
	AAYS	■	□	□		
	ABAT	■	□	□		
	ABCD	■	□	□		
	ACNR	■	□	□		
	ACNS	■	□	□		
	AZ07-7301	■	□	□		
	AZ07-7462	■	□	□		
	AZ07-7298	■	□	□		
	CO07-2003	■	□	□		
	CO07-2014	■	□	□		
	CO07-2015	■	□	□		
	CO07-3070	■	□	□		
	CO07-6570	■	□	□		
	CO07-6570-120	■	□	□		
	NM02-4452-human	■	□	□		
	NM02-4452-mouse	■	□	□		
	NM02-4476-306	■	□	□		
	NM02-4477-309	■	□	□		
	YPE	■	□	□		
	YPF	■	□	□		
	YPG	■	□	□		
	YPH	■	□	□		
	YPI	■	□	□		

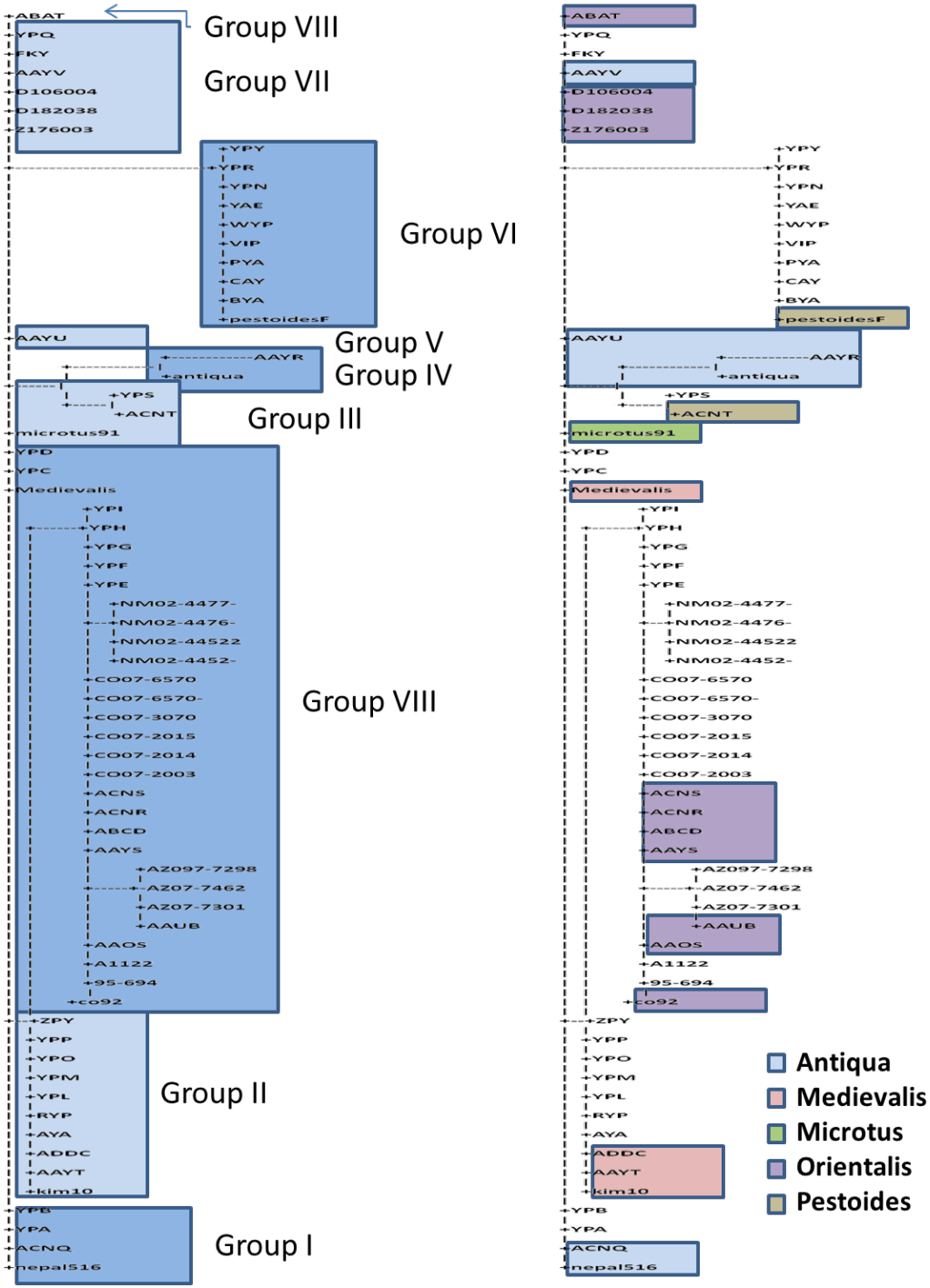

CRISPR 1
Strain
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
Group VII
Z176003
D106004
D182038
AAYV
FKY
YPQ
Group VIII
CO92
95-694
A1122
AAOS
AAUB
AAYS
ABAT
ABCD
ACNR
ACNS
AZ07-7301
AZ07-7462
AZ07-7298
CO07-2003
CO07-2014
CO07-2015
CO07-3070
CO07-6570
CO07-6570-120
NM02-4452-human
NM02-4452-mouse
NM02-4476-306
NM02-4477-309
YPE
YPF
YPG
YPH
YPI

PHYLOGENY OF *Y. PESTIS*: OUR STUDY



noblis
For the best of reasons

Evidence for Horizontal Gene transfer



noblis
For the best of reasons

Acknowledgments

- Significant contributions to this research project were made by:
  - Katharine Jennings, PhD
  - Mitchell Holland
  - Matthew McCoy

noblis

For the best of reasons